Infection, Genetics and Evolution 8 (2008) 593–602

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Rate heterogeneity in the evolution of *Helicobacter pylori* and the behavior of homoplastic sites

Richard J. Meinersmann [a,*], Judith Romero-Gallo [b], Martin J. Blaser [c]

[a] *USDA Agricultural Research Service, Athens, GA, United States*
[b] *Vanderbilt University, Nashville, TN, United States*
[c] *New York University School of Medicine and VA Medical Center, New York, NY, United States*

ARTICLE INFO

Article history:
Received 7 February 2008
Received in revised form 9 April 2008
Accepted 11 April 2008
Available online 18 April 2008

Keywords:
Homoplasy
Recombination
Transitions
Transversions
Helicobacter pylori

ABSTRACT

Helicobacter pylori are bacteria with substantial inter-strain variability and phylogenetic reconstructions of sequence data from the organism have common homoplastic sites. Although frequent recombination events have been proposed to contribute to the variation, the effects of nucleotide substitution rate heterogeneities on the reconstruction of *H. pylori* genealogies have not been studied. We analyzed the substitution pattern of a housekeeping gene, a homologue of the ribonuclease reductase gene (*rnr*), to characterize rate heterogeneities between 11 *H. pylori* isolates. Evidence of limited recombination was demonstrated by the Sawyer's runs test, but the homoplasy test and site-by-site compatibility tests indicated frequent recombination events. Within the 1935 nucleotide gene, 292 sites were polymorphic with an average pair-wise difference of 5.01%. Xia's distances for amino acids at non-synonymous codon substitution sites were smaller at homoplastic sites than at sites that were not homoplastic. Transitions were significantly more common among homoplastic than among non-homoplastic nucleotide substitutions. Simulations of evolution with or without recombination indicated the transition/transversion ratio is expected to be higher in homoplastic sites with no recombination. Despite evidence of recombination, analyses of the *rnr* genealogy does not show a random tree but rather base substitution behaviors characteristic of both recombination and substitution saturation at some sites. Analyses of sequences in the *H. pylori* multilocus sequence-typing database provided similar evidence for substitution saturation in multiple housekeeping genes.

Published by Elsevier B.V.

1. Introduction

Homoplasies are defined as when loci (sites in a sequence) have a number of steps on a reconstructed phylogenetic tree equal to or greater than the number of different alleles (characters) found at that locus (Maddison and Maddison, 1992). A homoplasy is an allele identified more than once in a phylogenic tree that is not derived from a common ancestor. Thus, a given character at a specific site will appear to arise more than once in the tree. Assuming an accurate reconstruction of the phylogeny, homoplasies can arise as a result of evolutionary convergence, parallelism or reversals. Analysis of multiple DNA sequences is the most informative method of population genetics, but accuracy is dependent on use of correct models of evolution (Hillis et al., 1994; Li, 1997). Rate heterogeneity, which is variation in substitution rate among the different nucleotides or base positions, when extreme, can lead to site saturation, i.e., multiple events at sites, some of which will show apparent reversion to an ancestral state (Xia, 2000b). However, horizontal gene transfer can also introduce homoplasies because genetic materials with different phylogenetic histories are brought together such that apparent reversals are introduced. Some researchers believe that horizontal gene transfer is more likely than reversals due to point mutations to introduce homoplasies in bacteria (Falush et al., 2001), but at this time there are no diagnostic methods to distinguish between these two mechanisms.

Helicobacter pylori plays an important role in gastroduodenal diseases in humans and has remarkable adaptations that allow persistent survival in the gastric niche (Doig et al., 1999; Montecucco and Rappuoli, 2001). *H. pylori* also are extremely variable, with a low probability that independent isolates will be the same (Go et al., 1996; Han et al., 2000) and appear panmictic

[*] Corresponding author at: USDA Agricultural Research Service, Russell Research Center, P.O. Box 5677, Athens, GA 30604, United States. Tel.: +1 706 546 3236; fax: +1 706 546 3633.
E-mail address: Rick.Meinersmann@ars.usda.gov (R.J. Meinersmann).

1567-1348/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.meegid.2008.04.003

(Salaun et al., 1998; Suerbaum et al., 1998), meaning that genetic exchange occurs too rapidly for linkage of alleles at different loci to be observed. Studies of the population genetics of *H. pylori* may be helpful in identifying selective pressures on the organism. Reconstructing ancestral relationships between strains is confounded by recombination events, as well as by rate heterogeneity. Although most of the limitations in *H. pylori* phylogenetics have been attributed to recombinant events (Maynard-Smith and Smith, 1998; Suerbaum et al., 1998), rate heterogeneity has not been fully considered.

This study began as a survey of *H. pylori* diversity within individual patients (Romero-Gallo and Blaser, unpublished). In one patient separate isolates displayed a gene with heterogeneity and sequence analysis showed high similarity to the gene encoding ribonucleotide reductase (*rnr*) (Tobe et al., 1992; Cheng et al., 1998), a housekeeping gene expected to be well conserved. Since sampling of partial *rnr* sequences from several independent isolates of *H. pylori* showed substantial diversity, the present study sought to determine the mechanisms underlying the observed patterns of *rnr* diversity, and whether units of recombination could be identified. Our findings indicated that there was a bias in the type of changes seen among homoplastic sites that was more consistent with rate heterogeneity than by horizontal gene transfer. We further tested to see if the bias was seen in controlled computer simulations of evolution and in other examples of *Helicobacter* gene evolution.

2. Materials and methods

2.1. Gene sequences

The *H. pylori* isolates studied are listed in Table 1. Isolates 97-793 and 97-645 were isolated from the fundus and antrum, respectively, from biopsies obtained simultaneously from the same patient, as described above. This pair of isolates appeared to be clonal variants, based on RAPD analysis (data not shown). DNA sequences for strains 26695 and J99 were obtained from the genomic sequences deposited in Genbank (accession numbers AE000630 and AE001544). For the remaining isolates, templates for DNA sequencing were generated by polymerase chain reaction using primers listed in Table 2. Sequencing was performed by dye terminator reactions with the same and other primers (Table 2) followed by analysis on an ABI Prism 373 automated sequencer. The quality of the sequences was evaluated and contiguous sequences were constructed with Sequencher (Gene Codes Corporation, Ann Arbor, MI).

Table 1
H. pylori isolates used in study

Strain designations		Source
97-679	A-4	India (Ladakh)
97-645	A-30[a]	India (Ladakh)
97-793	F-30[a]	India (Ladakh)
98-884	ATCC51407	Monkey
98-927	1308-3	Monkey
98-924	1309-3	Monkey
97-12	DB 011	Hong Kong
26695		UK (genomic sequence strain)
Hpk5		Japan
HPJ166		USA
J99		USA (genomic sequence strain)

[a] A-30 and F-30 were the index isolates from the stomach antrum and fundus, respectively, of the same person and appear to be clonal variants by RAPD (data not shown); both were included to bias toward finding recombination junctions.

2.2. Analyses

DNA sequences were aligned using ClustalX (Thompson et al., 1994). No gaps were introduced into the alignment and no additional editing was necessary. Phylogenetic reconstructions were performed using PAUP* ver. 4.0b4a (Swofford, 1998) and with DAMBE ver. 4.0.39 (Xia, 2000a). Window analyses of nucleotide diversity were produced using MULTICOMP ver. 1.01 (kindly provided by Ruiting Lan, University of Sydney, Sydney, New South Wales, Australia) using values for synonymous (K_s) and non-synonymous (K_a) substitution rates calculated by the method of Li (1993). Substitution patterns based on phylogenies were analyzed using DAMBE (Xia, 2000a). Homoplasies were identified on the same phylogenies using PAUP*. Homoplasies were observed at identical sites when the rooting was changed to other isolates or to midpoint rooting. Recombination analyses using site-by-site compatibility methods were done with RETICULATE (Jakobsen and Easteal, 1996) and with SITES (Hey and Wakeley, 1997). Maynard–Smith's homoplasy test was performed with HOMOPLASY (Maynard-Smith and Smith, 1998), and Sawyer's runs test was performed using GENECONV (Sawyer, 1999). The informative sites test was performed with PIST (Worobey, 2001) on datasets containing only the third codon position base. RECOMBINE (Kuhner et al., 2000) was run with trees generated in PAUP* as the initial phylogeny, the transition/transversion ratio set to 4.0, the base frequencies and initial value of theta determined by the program, and the remainder of the parameters set at the default values. The shape parameter (α) of the gamma distribution was determined with the data generated by PAUP* using the HKY85 model and four rate categories in a maximum likelihood test (Swofford, 1998). MODELTEST ver. 3.7 (Posada and Crandall, 1998) was used to compare different models of evolution.

2.3. Computer simulations

Computer simulations were run with "Seq-Gen" (Rambaut and Grassly, 1997) in which either recombination or rate heterogeneity without recombination was responsible for the bulk of the homoplasies in the data. The HKY85 model of evolution (Hasegawa et al., 1985) was used with the base composition and the transition/transversion ratio set to approximate those calculated from the *rnr* data. Seq-Gen uses guide trees to generate simulated evolutions. For non-recombinant data, a single guide tree was used (the tree reconstructed from the *rnr* sequence) and the rate heterogeneity shape parameter (α) was set to the value calculated for the *rnr* data. To simulate recombination the rate heterogeneity was left at the default (equal rates for all sites) and the sequence was divided into nine partitions. The nine partitions were defined by bootscan analysis (Salminen et al., 1995) implemented in RDP (Martin and Rybicki, 2000) of the *rnr* data and trees that were used as guide trees by Seq-Gen were deduced for each partition with PAUP*. One thousand simulations were run with either high rate heterogeneity (recombination essentially nil) or with high recombination (rate heterogeneity essentially nil). The output for each simulation was put through PAUP* to diagnose changes at each site to determine if they were homoplastic and if they were transitions or transversions. Scripts were developed in SAS and Excel to tabulate the number of changes in each simulation for each category: homoplastic or non-homoplastic, and transition or transversion. The informative sites among the non-homoplastic sites were also considered separately since all homoplastic sites are automatically informative. Excel was used to tabulate the frequency of each observation and to produce the graphics.

Table 2
Oligonucleotide primers used to generate *rnr* sequence data

Gene	Orientation	Primer designation	$5' \rightarrow 3'$ sequence	Positions[a]
aroE(*HP1249*)	Forward	FB	GAGCGGGATCCGATTTGGCGTATGGGTTTTTAA	1324522-1324543[b]
rnr	Forward	F1	CCTTACTAGAATTCGCCAGG	427-446
rnr	Forward	F2	TTTGGGCGTGGTTTTAGAGG	1740-1759
rnr	Forward	F3	CTTACCCCCATCTTTTAAAAC	222-242
rnr	Forward	F4	TAAAGCGAGCGATTTTAAAGA	609-629
rnr	Forward	F5	TTTGCAACAAAGCCTTTTAGG	1035-1055
rnr	Forward	F6	AAATTTAGCCCTTTATAGCCC	1455-1475
rnr	Forward	F7	AAAGGTGCGCGTTACAATCAC	1845-1865
rnr	Reverse	R1	GGTGATAAAAGGGATGTGAG	663-644
rnr	Reverse	R2	CCTTTCTGTGATTTCGCCTC	1894-1914
rnr	Reverse	R3	TAAGCGAGCGAATTTGCGCTT	1698-1718
rnr	Reverse	R4	TAAAGGCGTTTTTGCTGCTCC	1323-1302
rnr	Reverse	R5	TTCATACACTAAAGCCAGGCG	912-892
rnr	Reverse	R6	GGGTCTTCTAAAGCCCCTAAA	515-495
rnr	Reverse	R7	TTGCCTATATCAAAGCCTTCT	147-126
rnr	Reverse	NR2	GCTCTTTAATCCTTTCTGTGATTTCGCCTC	1924-1895
HP 1247	Reverse	RB	GAGCGGAATTCATCTGGCTTTTTTCATAATCGC	1322593-1322572[b]

[a] Positions are based on the start of *rnr* unless indicated by[b], which are the positions based on the annotation of the total genomic sequence of *H. pylori* strain 26695.

2.4. Large data sets

Sequences were obtained from the *H. pylori* multilocus sequence-typing (MLST) database (http://pubmlst.org/helicobacter/) (Jolley et al., 2004). Only loci that could be aligned without gaps (*atpA*, *efb*, *mutY*, *ppa*, *trpC* and *ureI*) were used. Phylogenies were constructed for each locus using PAUP* using HKY85 distances. The number of changes at each site, expressed as the consistency index ($CI = m_i/s_i$, where s_i is the number of number of reconstructed steps for character i and m_i is the maximum numbers of steps that might have occurred at site i without any homoplasies (Maddison and Maddison, 1992)) and the transition or transversion characteristic of each change was determined. Since the CI for nucleic acid data is a progression of 1, 2, or 3, divided by the number of changes, the percent transversions was plotted against the inverse of the CI and regression analysis was performed using Statistica (StatSoft, Inc., Tulsa, OK).

3. Results

3.1. Analysis of recombination in H. pylori rnr

Recombination affects the accuracy of phylogenetic reconstruction (Maynard-Smith and Smith, 1998). To determine parameters to include in phylogenetic reconstructions of *H. pylori rnr*, the sequences of the whole gene (1935 bp) from 11 isolates were evaluated for recombination after alignment with ClustalX revealed 292 polymorphic sites. Analysis with RETICULATE showed many polymorphic sites with apparently distinct phylogenetic histories (neighbor similarity score = 0.641, P value < 0.001), indicating multiple possible recombination events (see Fig. 1A). A similar site-by-site congruencies test, generated with the program SITES, indicated multiple recombinations and pointed to the possibility of 48 recombination intervals. Maynard–Smith's homoplasy test (Maynard-Smith and Smith, 1998) showed high homoplasy excess

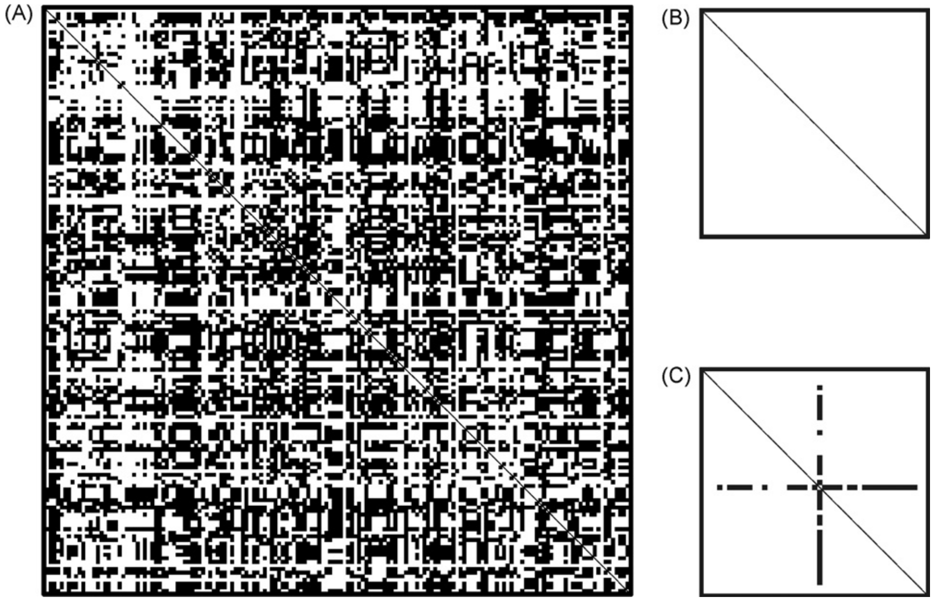

Fig. 1. Results matrixes of reticulate analysis (as implemented in RETICULATE) for (A) the observed dataset of the 11 *H. pylori rnr* genes with all homoplasies intact, (B) the dataset with all identified homoplastic sites removed, and (C) an example in which one of the homoplastic sites was restored. The graphic displays polymorphic sites aligned on the vertical and horizontal axes. All the sites underwent pair-wise comparisons for compatibility. A black mark is displayed at the intersection for two sites if the changes cannot be explained by clonal evolution.

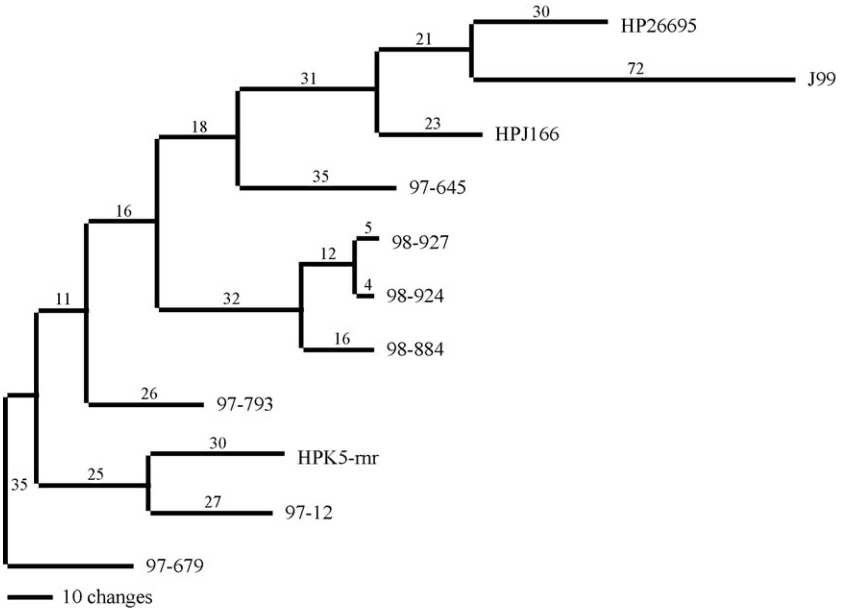

Fig. 2. Phylogram reconstructed by parsimony analysis of the DNA sequences of *rnr* from 11 isolates of *H. pylori*. The most parsimonious tree is shown (only one most parsimonious tree was found; other methods gave the same branch pattern; see text). The numbers along the lines indicate the number of base changes from node to node or from node to the terminal leaf of each branch. The tree is rooted on 97-679, which had the sequence most similar to that seen from the *C. jejuni rnr* homologue (Cj0631c).

(ratio = 0.658), also consistent with frequent recombination. Sawyer's runs test, implemented in GENECONV (Sawyer, 1999), identified a possible recombination event between isolate 98-884 and a recent ancestor of isolates 98-924 and 98-927. A breakpoint in the sequence was identified at about base 706 with the bases between 1 and 706 common in all three isolates and numerous polymorphisms occurring in the remainder of the sequence. In total, these observations suggest a history of extensive recombination in *rnr*, consistent with prior observations of *H. pylori* genes (Salaun et al., 1998; Suerbaum et al., 1998; Falush et al., 2001).

3.2. Phylogenetic reconstruction of rnr sequences

Another phenomenon that could affect assessment of recombination in the history of *rnr* is the occurrence of mutations so frequently that multiple mutations occur at individual sites, a condition known as "site saturation" (Xia, 2000b). To investigate this possibility, we began by performing phylogenetic reconstruction of *rnr* DNA sequences. Initially, phylogenetic reconstruction of the *rnr* DNA sequences was performed using the *C. jejuni rnr*-homologue as a root. The codon bias of *C. jejuni* is very different from that for *H. pylori* (see codon bias tables at http://www.kazusa.or.jp/codon/), which will not affect phylogenetic reconstructions but may introduce biases into our analyses of the types of nucleotide changes. Therefore, all subsequent analyses of patterns of nucleotide changes and phylogenetic reconstructions were performed without the *C. jejuni* sequence. Instead, the sequence of isolate 97-679 was used as the root because it was closest to the root when the *C. jejuni* sequence was included. Reconstruction by maximum parsimony using an exhaustive search in which every possible tree was tested in PAUP* yielded a single optimal (shortest) tree (Fig. 2).

Each of the distance-based methods used [TN93 for nucleotides (Tamura and Nei, 1993), Poisson proportion for amino acid sequence (Xia, 2000a), maximum likelihood (Yang et al., 1995), HKY85 (Hasegawa et al., 1985) and Li's codon-based distances weighted equally for synonymous and non-synonymous substitutions (Li, 1993; Xia, 2000a) gave identical branching patterns with slight differences in branch lengths. In addition, MODELTEST (Posada and Crandall, 1998) revealed that the HKY85 model with invariant sites and rate heterogeneity best fit the data. Using these parameters also yielded a tree with identical branching pattern to the parsimony tree illustrated (Fig. 2). These results show that each of the common methods for phylogenetic reconstruction are equally valid for studying nucleotide substitution patterns in *rnr* and that the tree-based substitution patterns discussed below would not be different if the model used for the analysis was changed.

3.3. Non-synonymous substitutions in rnr sequences

A prediction based on the Neutral Theory of Evolution is that the frequency of non-synonymous substitutions will vary directly with the degree of similarity of the substituted amino acid (Kimura,

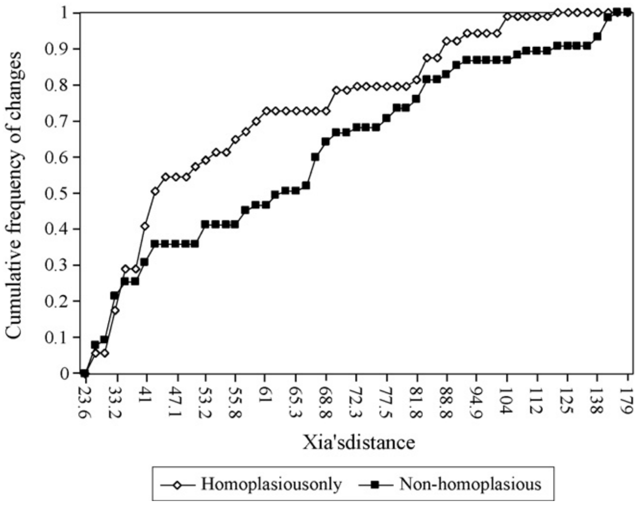

Fig. 3. Cumulative frequency of non-synonymous changes relative to Xia's distance between the amino acids in each change based on the reconstructed phylogram shown in Fig. 2. The steeper increase for the homoplastic only sites (open diamonds) relative to the non-homoplastic sites (closed boxes) is consistent with a preference for smaller changes at homoplastic sites (for the difference of the means, $P = 0.002$).

1968; Xia, 2000b). Amino acids that are dissimilar are more likely to adversely affect the function of the protein and will be selected against (Nei, 2005). The degree of dissimilarity can be expressed as Xia's distance, a measure of the differences between amino acids, which can range from 23.6 to 256. Using the reconstructed phylogram (Fig. 2), the putative ancestral states and all the necessary changes to derive the present sequences were derived by a maximum likelihood analysis. This revealed 202 non-synonymous substitutions, which had a mean for all amino acid substitution distance of 62.6, with a frequency skew toward changes involving smaller distances (data not shown). Thus, as expected, non-synonymous changes that resulted in minor changes in protein conformation have been more tolerated.

A further expectation drawn from the Neutral Theory of Evolution is that substitutions with smaller differences are more likely to be involved in multiple events (Kimura, 1968; Xia, 2000b), i.e., the same-site may have the same change more than once. In such a case, analysis of an evolutionary tree is more likely to indicate a homoplasy at that site. Of the 202 non-synonymous substitutions in the *rnr* phylogeny (Fig. 2), 114 were identified as homoplastic by parsimony analysis implemented in PAUP (Swofford, 1998). Using Xia's metric for amino acid distances, viewing the cumulative frequency of changes relative to distance of amino acid change (Fig. 3) the homoplastic changes have a median distance of less than 45 whereas the median for non-homoplastic changes is about 65. The mean of the distances in the homoplastic changes was 54.4, which was significantly lower (*t*-test for independent variables $P = 0.002$) than the non-homoplastic changes with a mean distance of 68.3. This difference is consistent with homoplastic sites arising and persisting because of tolerance of changes with small amino acid differences.

3.4. Nucleotide substitution heterogeneity in rnr sequences

The average pair-wise difference of the *rnr* nucleotide sequences in this study was 5.01%, with a maximum difference of 7.60%. Using the deduced *rnr* evolutionary tree (Fig. 2) to catalogue the nucleotide substitutions from the root to each extant sequence showed that synonymous changes exceeded non-synonymous changes, as expected (Tables 3 and 4) and the transition/transversion (Ts/Tv) ratios correlated with the codon position pattern. Since transversions at third codon position are more likely to be non-synonymous than transitions, transitions exceeding transversions is expected. All independent changes at the second base position are non-synonymous and transitions and transversions would have a similar effect on amino acid substitution distances. Similarly, in a random sequence, independent changes in the first codon position are equally likely to be synonymous regardless of whether the change is a transition or transversion. Therefore, the Ts/Tv ratio at the first and second codon positions does not reflect selection based on the structural requirements of the protein product, but must reflect mutational preferences for each position. In *H. pylori rnr*, this is reflected by the low prevalence of G and C in the second codon position (Table 5), indicating that substitution-site bias is at least partially due to codon preferences. Results for *rnr* from strain 26695 (Table 5) are nearly identical for all other *H. pylori* strains tested, including J99 (data not shown).

Table 3
Location of substitutions by position of the changes in *rnr*[a]

Category of substitution	Site of differences[b]	Frequency
None (identical)		11,815
Synonymous	100	12
	010	0
	001	221
	110	0
	101	5
	011	0
	111	0
Subtotal		238
Non-synonymous	100	116
	010	50
	001	12
	110	4
	101	9
	011	10
	111	1
Subtotal		202

[a] Based on tree shown in Fig. 2.

[b] The number indicates the three positions of the codon, 0 indicates no change at that position and 1 indicates a change.

Table 4
rnr substitution patterns, by codon position[a]

Codon position	Nucleotide change				Transition/ transversion ratio
First					
↓ from/to →	A	G	C	T	
A	–	28	8	1	
G	45	–	7	4	
C	5	3	–	24	
T	4	1	17	–	
Total				147	3.45
Second					
↓ from/to →	A	G	C	T	
A	–	10	5	2	
G	13	–	0	2	
C	5	4	–	10	
T	1	4	9	–	
Total				65	1.82
Third					
↓ from/to →	A	G	C	T	
A	–	50	7	3	
G	47	–	7	10	
C	3	5	–	56	
T	4	6	60	–	
Total				258	4.73

[a] Based on the phylogeny shown in Fig. 2.

Of the 469 base changes observed in the *rnr* phylogeny (Fig. 2), 286 were at homoplastic sites, of which 230 were transitions (Ts/Tv = 4.11). Of the remaining 47 non-homoplastic base changes at informative sites, 26 were transitions (Ts/Tv = 1.24). Thus, homoplastic base changes had significant over-representation of transitions (Chi-square, $P < 0.01$). There are greater than 34 million possible unrooted trees that will fully resolve 11 isolates such as used in the present analysis. Only sites that are parsimoniously informative can be homoplastic, and most of the possible trees will have all of the informative sites indicated as a homoplastic change. Therefore, most randomly selected trees will

Table 5
Frequency of *rnr* nucleotide usage in *H. pylori* strain 26695

Codon position	Frequency per site[a]			
	T	C	A	G
1	0.25	0.16	0.30	0.29
2	0.33	0.17	0.37	0.13
3	0.31	0.19	0.30	0.20
All positions	0.297	0.170	0.326	0.208

[a] Based on singleton nucleotide usage.

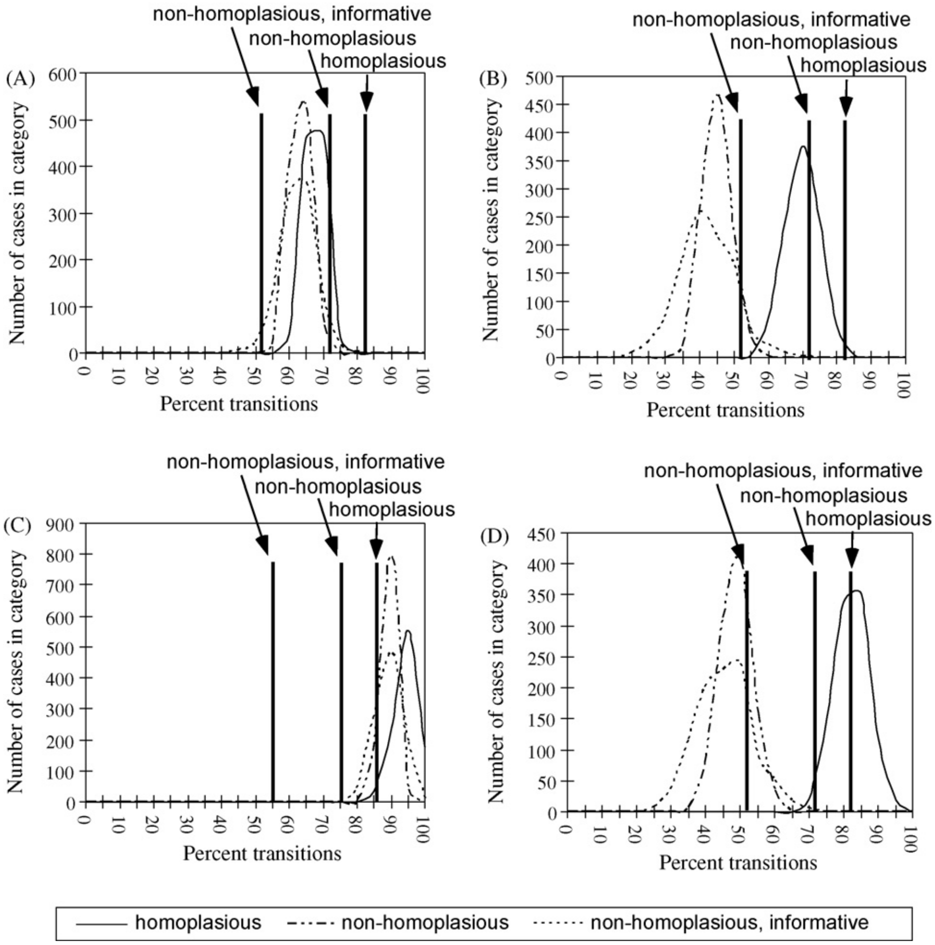

Fig. 4. Frequency of transitions in simulations. Sequence evolution simulations were performed with four different conditions: (A) and (B) had Ts/Tv ratio set at 4.0; (C) and (D) had Ts/Tv ratio set at 8.0; (A) and (C) used data partitioning to simulate recombination with the rate heterogeneity parameter (α) set to the default (essentially no rate heterogeneity); (B) and (D) used a single tree (no recombination) with the rate heterogeneity parameter (α) set to 0.014. Each graph shows the distribution of percent of changes that were transitions for all non-homoplastic sites (irregular dashed line), homoplastic sites (solid line), and informative homoplastic sites (regular dashed line). Each graph also has vertical bars that indicate percent of transitions for all the non-homoplastic sites, the homoplastic sites, and the informative homoplastic sites in the *rnr* data. The ordinate shows the absolute number of occurrences of the transition level shown on the abscissa (in bins of 5%) within 1000 simulations.

include that all of the informative sites are homoplasies, and the average number of transitions and transversions at homoplastic sites will equal all the possible transitions and transversions at the informative sites. A significant departure from the proportion of total possible transitions and transversions, such as seen in the *rnr* data, is not expected on this basis. Intuitively, for homoplasies that arise by recombination, there is no mechanism to favor transitions over transversions.

3.5. Homoplastic site behavior

To study the behavior of homoplastic sites, sequences were generated to simulate evolution under conditions that would produce homoplasies due to either rate heterogeneity that resulted in site saturation or due to recombination. Using Seq-Gen, 1000 simulations were generated with parameters derived from the *rnr* data: model = HKY85, sequence lengths = 1935 bp, transition/transversion ratio = 4.0, base frequency ratios—A = 32%, C = 17%, G = 21%, T = 30%. For homoplasies without recombination, a single tree was used as the guide (no-partitions) and the rate heterogeneity shape parameter (α) was set to 0.0148, the value determined from the *rnr* data. Simulations to produce homoplasies due to recombination were done with nine partitions, the same size as estimated for the *rnr* data, with separate guide trees for each partition, and alpha was set to the program's default (virtually no rate heterogeneity). The frequencies of transitions and transversions were tabulated for homoplastic and non-homoplastic sites for each condition (with and without recombination) (Fig. 4A and B). The Ts/Tv ratio was usually higher in evolution with recombination (simulations with partitions) in both homoplastic and non-homoplastic sites, but lower in the latter. The *rnr* data do not have strong membership in the clustering produced by any simulated condition, but the simulation generator apparently averaged the Ts/Tv ratio between the homoplastic or non-homoplastic sites, whereas the *rnr* data appear to have independent categories. Re-running the simulations with the same parameters except setting the Ts/Tv ratio to 8.0 (Fig. 4C and D), yielded results consistent with the *rnr* data. The smaller Ts/Tv ratio in the *rnr* data for the informative non-homoplastic sites versus all non-homoplastic sites may reflect that informative sites may be older changes, subject to selective pressures for a longer time. The simulators do not have selective pressure parameters.

If homoplasies are introduced by recombination, neighboring sites (i.e., one polymorphic site and the next one or ones in a sliding window) would more likely be homoplastic, due to being transferred together in a recombinant event. In the *rnr* data set, there were 30 instances in which $\geq$2 neighboring polymorphic sites were homoplastic. Since 17 of these neighbor groups involved

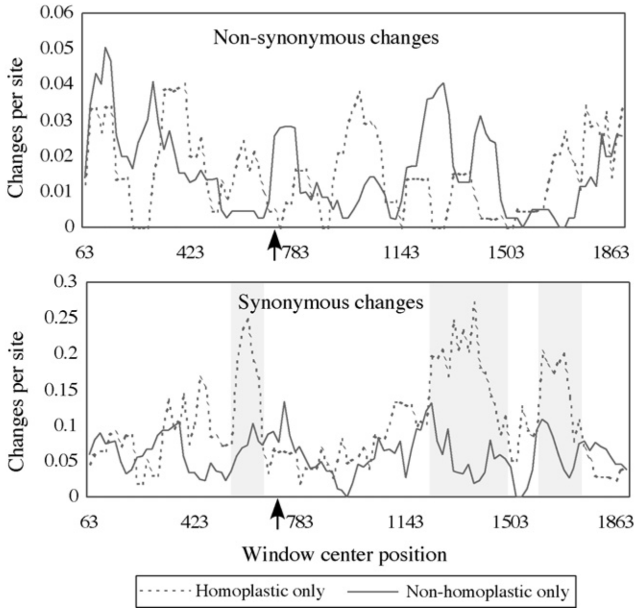

Fig. 5. Window analysis of substitution rates. Substitution rates were calculated for windows of 120 bp with jumps of 3 bp over the length of the aligned *rnr* sequences from 11 strains. The analyses were repeated with only homoplastic sites included or only non-homoplastic sites included. The rates for the synonymous and non-synonymous substitutions are shown for each category. The shaded areas are regions in which synonymous changes are substantially higher in homoplastic sites, possibly indicating non-Darwinian selection. The arrows indicate the location of a possible recombination junction identified by Sawyer's test.

different branches of the evolutionary tree, this was not consistent with clustering due to incorporation of sequence blocks by recombination. Of the other 13 neighbor groups, 7 were within single codons, and the only homoplastic neighbors involving the same branches that were more than 3 bp apart were two that were 6 bp apart. Thus, the neighboring homoplasies observed are consistent with either small recombination events or with substitution pressures due to nucleotide neighbor (e.g., codon) preferences.

3.6. Gene site substitution rate heterogeneity

We next sought to determine whether there was substitution rate heterogeneity by inspecting the density of putative substitutions along subsections of the sequence. Among the 11 sequences of the 1935 bp *rnr*, the 292 polymorphic nucleotide sites observed yielded 126 polymorphic deduced amino acids. Using a 120 bp window, the synonymous and non-synonymous substitution rates were highly variable, consistent with differing selective pressures (Fig. 5). Based on the Neutral Theory of Evolution (Kimura, 1968), non-synonymous homoplasies should be over-represented at positions that could tolerate changes in the protein structure, and synonymous substitution rates should be similar for both homoplastic and non-homoplastic sites. However, in three regions the synonymous changes among homoplastic sites were substantially higher than among the non-homoplastic sites (Fig. 5). This observation provides evidence that selection is active at the DNA level, not only the protein sequence level, in driving the homoplastic changes. Since homoplasies due to recombination events introduce parallel changes in synonymous and non-synonymous substitution rates, recombination events alone cannot explain the observations presented in Fig. 5. In addition, there is no apparent correlation of the recombination breakpoint at base 706 identified with the Sawyer's runs test (see above) with the synonymous and non-synonymous substitution rates.

A more formal measure of rate heterogeneity is the determination of the shape parameter (α) of the gamma distribution (Swofford et al., 1996). The determined shape parameter for the *rnr* data was 0.0148, characteristic of a high degree of rate heterogeneity (Swofford et al., 1996). When the homoplastic sites were removed, the rate heterogeneity resulted in an estimated shape parameter of infinity, indicating that all the sites have an equal rate of change (Swofford et al., 1996). This comparison indicates that the sites identified as homoplastic create the signal for rate heterogeneity.

3.7. Comparisons of recombination parameters with and without homoplasies removed

Since the homoplastic sites appear to drive the measure of rate heterogeneity, we next performed tests of recombination that are independent of the tree structure, with and without the homoplasies removed. Repeating the RETICULATE analysis with homoplasies removed, showed no evidence of recombination (Fig. 1B). Restoring any of the homoplastic sites also restored a signal for recombination (e.g. Fig. 1C). When the SITES test of congruency (Hey and Wakeley, 1997) was performed with the dataset that had the homoplasies removed, exactly the same dichotomy of results was observed. Using the Metropolis–Hastings Markov Chain Monte Carlo genealogy sample in RECOMBINE (Kuhner et al., 2000), the recombination rate (rho) was estimated to be 6.6×10^{-5} with all data analyzed; the rate was reduced to 1×10^{-7} when homoplasies were removed.

Another way of measuring recombination frequency is the informative sites test (Worobey, 2001), implemented in the program PIST. Briefly, the test creates a "*Q* score" by comparing the numbers of observed two-state informative sites (sites with two of the possible four states (A, C, G, and T), and in which both states occur in more than one sequence each) with the numbers in simulated data sets generated under the constraints of a clonal model of evolution that follows the derived tree (Fig. 2). The significance of the departure from clonality is determined by repeated simulation of the phylogeny, and then calculating the frequency that the observed two-state informative sites exceed the simulated data. PIST was run with 1000 replicates with the observed dataset with or without the homoplasies intact. The *Q* score for the intact observed data was 0.493, which was exceeded by only 2 of the simulations ($P < 0.003$). The *Q* score for the dataset with the homoplastic sites removed was 0.208, which was exceeded by that for 995 simulations ($P < 0.996$). These results provide additional evidence that identification of the homoplastic sites also identifies the sites that impart recombination signal.

3.8. Analysis of MLST gene sequences from H. pylori

It is reasonable to expect that other genes within *H. pylori* will have similar phylogenetic histories. A readily available test set was derived from the MLST database (http://pubmlst.org/helicobacter/). The alignments for *efb*, *atpA*, *mutY*, *ppa*, *trpC*, and *ureI* which range in length from 398 bp to 627 bp and had from 224 to 361 (mean 323) taxa in each alignment, were prepared and analyzed. This resulted in a total of 2896 characters of which 1038 were polymorphic. Only 43 of these sites were informative but not homoplastic. Adding taxa to the alignment increased the proportion of sites that were homoplastic, with the number of remaining non-homoplastic sites too small for meaningful statistical analysis.

Sites that are more permissive to change, for instance because the change would be synonymous or would result in a small amino acid change, can be predicted to more likely be homoplastic. It follows that if the homoplasies are due to mutations rather than

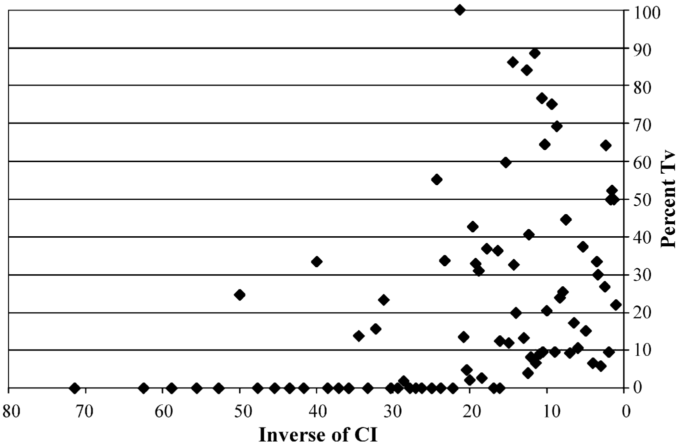

Fig. 6. Percent of transversions at different levels of the consistency index (CI) for pooled MLST data from 6 genes, representing a cumulative 2896 bp. The CIs are transformed to their inverses and plotted in reverse order so that smaller CIs are to the left. The ordinate shows the percent of transversions at each CI.

recombination, then the more homoplastic sites would have a greater over-representation of transitions since transitions are favored in site mutations. Once an evolutionary tree is reconstructed for an alignment the consistency index for each polymorphic site can be calculated. Of the 1038 polymorphic sites in the MLST test set, there were 81 different Consistency Indices (CIs) (see Section 2) ranging from 0.014 to 1.0. There were a total of 10,881 changes in the tree, including 9327 transitions and 1554 transversions (Ts/Tv = 6.0). The inverse of the CI was plotted against the percent Tv (readily related to Ts and Ts/Tv ratio) (Fig. 6), to correlate CI with the transition/transversion ratio. Regression analysis (Statistica) yielded $R^2 = 0.177$ with $P = 0.00094$, thus we can be highly confident that the CI explains some proportion of Tv. There often appeared to be a "jackpot"-like phenomenon in which all of the changes at a site were identical, probably resulting in some sites that were not representative, possibly accounting for much of the variation. Recombination certainly accounts for some of the variation as well. However, sites with lower CIs most likely have all transitions (Fig. 6).

4. Discussion

For the genes studied, the behavior of homoplastic sites can be explained by either recombination or site saturation with strong selection for maintenance of function. We did not anticipate the observed difference in transition/transversion ratio between homoplastic and non-homoplastic sites. Since we could not locate prior studies of transition/transversion ratio of homoplastic sites, computer simulations of phylogenies creating homoplasies with and without recombination were performed. The simulations clearly show that recombination alone cannot explain the findings.

H. pylori is a highly diverse species (Go et al., 1996; Han et al., 2000); understanding the mechanisms responsible for this diversity will help to elucidate its population ecology. For example, reassortive recombination requires association of multiple clones, minimally a donor and an acceptor, whereas mutations should be independent events occurring irrespective of the presence of other clones. With a panmictic population structure (Salaun et al., 1998; Suerbaum et al., 1998), there must have been frequent opportunities to exchange DNA between *H. pylori* strains. The only natural reservoir for *H. pylori* is the stomach of humans and other primates (Dunn et al., 1997). Humans usually acquire the organism in early childhood and remain carriers for life or until antimicrobial treatment (Taylor and Parsonnet, 1995). The family is the major unit of *H. pylori* transmission (Rothenbacher et al., 1999; Han et al., 2000; Raymond et al., 2004). Multiple subtypes of the organism have been identified in individual hosts (Hirschl et al., 1994; Taylor et al., 1995; Jorgensen et al., 1996; Morales-Espinosa et al., 1999; Kuipers et al., 2000), representing both different clones and diversification of the originally colonizing clone (Falush et al., 2001).

It has been well documented that diversification of an *H. pylori* strain in an individual human subject can occur by loss and/or addition of *H. pylori* genes (Israel et al., 2001). Falush et al. (2001) estimated the median size of imported DNA in a recombination event in *H. pylori* is about 417 base-pairs. Since *H. pylori* cells are naturally competent for transformation by naked DNA (Haas et al., 1993; Tsuda et al., 1993; Wang et al., 1993; Hofreuter et al., 2000; Israel et al., 2000), it is possible that free DNA from other clones is frequently incorporated into the genome of the colonizing strains. The Sawyer's runs test, as illustrated above with our data, provides strong evidence that reassortive recombination occurs in the phylogeny of *H. pylori*. Results of studies using hybridization arrays (Israel et al., 2001; Salama et al., 2000) indicate that insertions and deletions are relatively common events, and involve segments that include whole genes or groups of genes.

However, given the analyses presented here, the methods of inferring recombination events based on evolutionary consistency (i.e., homoplasy test, reticulate test, and site congruency test) are not completely reliable since these methods assume that most homoplasies are due to recombination events. Returning any single homoplastic site to the dataset without homoplasies restores several incongruent pairs of sites identified by reticulate analysis (for example, see Fig. 1C), which indicates that the identified homoplastic sites constitute the minimum set yielding signal that can be interpreted as evidence for recombination. The non-synonymous homoplastic nucleotide substitutions we observed had a smaller average Xia's distance of amino acid changes than did non-synonymous non-homoplastic nucleotide substitutions, which could be due to purifying selection after the mutation events, with subsequent unavailability for future recombination. However, this would not explain their nonrandom occurrence with respect to the transition/transversion ratio of nucleotide substitutions. Contrary to the expectation of random recombinant exchange (which would result in data leading to an erroneous phylogram and thereby identification of an excess of random homoplastic sites), we found that homoplastic nucleotide substitutions favor transitions over transversions at a higher rate than non-homoplastic substitutions. This only can be explained by mutation substitution rate heterogeneity with site saturation. Worobey's informative sites test (Worobey, 2001) is considered to take into account observed apparent rate heterogeneity; however, the test is dependent on the bases being permitted, within the constraints of the transition/transversion parameters, to change to any of the other three bases. Considering the strong codon bias observed with *H. pylori* (http://www.kazusa.or.jp/codon/), this assumption is probably not met and, thus, the test may not be sufficiently accurate.

The mutation rate in wild-type *H. pylori* isolates can be very high, sometimes exceeding the rate seen in *Enterobacteriaceae* mismatch-repair defective strains (Bjorkholm et al., 2001). Such high rates create conditions in which site saturation could influence interpretation of phylogenies, since both synonymous and nearly neutral polymorphisms exert little pressure for conservation. Since codon usage is a selective force, the substitution saturation rate is lower than otherwise expected, because some trinucleotide patterns are disfavored, which would reduce the pool of eligible substitutions. As far as we can determine, tools are not available to predict saturation rates in the presence of strong codon bias. Codon preference appears to be one driver of site rate heterogeneity, but since we have observed that rates of substitutions are not uniform over the length of *rnr*, at least one other driver of site rate heterogeneity may exist.

We confirm previous studies that inter-clonal recombination occurs in *H. pylori* (Go et al., 1996; Salaun et al., 1998; Suerbaum et al., 1998; Han et al., 2000; Falush et al., 2001), but at a rate that cannot be properly estimated with the phylogenetic models that have been previously applied. It can be argued that in the face of evidence of recombination, a phylogenetic reconstruction is meaningless. If that is the case, our analyses call into question several methods of determining recombination, at least with respect to quantifying recombination. This is because identifying homoplastic sites is entirely tree-dependent. We have shown that removing homoplastic sites from analyses removed the evidence of recombination that could be shown by site congruency tests and, thus, these tests give exactly the same results as analyses of tree-based analyses.

The alternative interpretation is that at least some valid information can be found in the reconstructed tree. We have shown that the homoplastic sites are not randomly partitioned. The biases, over-representation of transitions among homoplastic sites and over-representation of more slight amino acid changes among homoplastic non-synonymous sites, are consistent with multiple same-site point mutations. This is a signature of site saturation. The bias for over-representation for transitions was also shown for other genes from *H. pylori* with data from many more taxonomical units. The more changes that occur at a site, the more likely the changes are transitions, which in the simulation only occurred with mutations and not with recombination. It can be argued that the accuracy of the determination of homoplastic sites is dependent on the accuracy of the reconstructed phylogeny. Two factors will result in trees that are not accurate—the result of horizontal gene transfer (recombination) and the large number of taxa results in more errors in the reconstruction algorithms. However, the algorithms are likely to produce trees that are much better than random and associations found in turn to be non-random from analysis of any non-random tree is likely have a real component. Bootstrap trees may increase the accuracy but they also introduce more polytomies.

Our analyses suggest that the rate of recombination is likely to be much lower than currently estimated, which also is consistent with studies showing restriction barriers between *H. pylori* strains (Ando et al., 2000; Donahue et al., 2000; Xu et al., 2000). If all the homoplasies were due to rate heterogeneity, which is unlikely, the data from the RECOMBINE analysis (Kuhner et al., 2000) with and without the homoplastic sites included suggests that the recombination rate may be overestimated by more than 600-fold. Our analyses suggest that the relative contributions of recombination and point mutation to *H. pylori* diversity must be reappraised.

References

Ando, T., Xu, Q., Torres, M., Kusugami, K., Israel, D.A., Blaser, M.J., 2000. Restriction–modification system differences in *Helicobacter pylori* are a barrier to interstrain plasmid transfer. Mol. Microbiol. 37, 1052–1065.

Bjorkholm, B., Sjolund, M., Falk, P.G., Berg, O.G., Engstrand, L., Andersson, D.I., 2001. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. Proc. Natl. Acad. Sci. U.S.A. 98, 14607–14612.

Cheng, Z.F., Zuo, Y., Li, Z., Rudd, K.E., Deutscher, M.P., 1998. The *vacB* gene required for virulence in *Shigella flexneri* and *Escherichia coli* encodes the exoribonuclease RNase R. J. Biol. Chem. 273, 14077–14080.

Doig, P., de Jonge, B.L., Alm, R.A., Brown, E.D., Uria-Nickelsen, M., Noonan, B., Mills, S.D., Tummino, P., Carmel, G., Guild, B.C., Moir, D.T., Vovis, G.F., Trust, T.J., 1999. *Helicobacter pylori* physiology predicted from genomic comparison of two strains. Microbiol. Mol. Biol. Rev. 63, 675–707.

Donahue, J.P., Israel, D.A., Peek, R.M., Blaser, M.J., Miller, G.G., 2000. Overcoming the restriction barrier to plasmid transformation of *Helicobacter pylori*. Mol. Microbiol. 37, 1066–1074.

Dunn, B.E., Cohen, H., Blaser, M.J., 1997. *Helicobacter pylori*. Clin. Micro Rev. 10, 720–741.

Falush, D., Kraft, C., Taylor, N.S., Correa, P., Fox, J.G., Achtman, M., Suerbaum, S., 2001. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. U.S.A. 98, 15056–15061.

Go, M.F., Kapur, V., Graham, D.Y., Musser, J.M., 1996. Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. J. Bacteriol. 178, 3934–3938.

Han, S.-R., Zschausch, H.-C.E., Meyer, H.-G.W., Schneider, T., Loos, M., Bhakdi, S., Maeurer, M.J., 2000. *Helicobacter pylor*i: clonal population structure and restricted transmission within families revealed by molecular typing. J. Clin. Microbiol. 38, 3646–3651.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21, 160–174.

Haas, R., Meyer, T.F., van Putten, J.P., 1993. Aflagellated mutants of *Helicobacter pylori* generated by genetic transformation of naturally competent strains using transposon shuttle mutagenesis. Mol. Microbiol. 8, 753–760.

Hey, J., Wakeley, J., 1997. A coalescent estimator of the population recombination rate. Genetics 145, 833–846.

Hillis, D.M., Huelsenbeck, J.P., Cunningham, C.W., 1994. Application and accuracy of molecular phylogenies. Science 264, 671–677.

Hirschl, A.M., Richter, M., Makristathis, A., Pruckl, P.M., Willinger, B., Schutze, K., Rotter, M.L., 1994. Single and multiple strain colonization in patients with *Helicobacter pylori*-associated gastritis: detection by macrorestriction DNA analysis. J. Infect. Dis. 170, 473–475.

Hofreuter, D., Odenbreit, S., Puls, J., Schwan, D., Haas, R., 2000. Genetic competence in *Helicobacter pylori*: mechanisms and biological implications. Res. Microbiol. 151, 487–491.

Israel, D.A., Lou, A.S., Blaser, M.J., 2000. Characteristics of *Helicobacter pylori* natural transformation. FEMS Microbiol. Lett. 186, 275–280.

Israel, D.A., Salama, N., Krishna, U., Rieger, U.M., Atherton, J.C., Falkow, S., Peek, R.M., 2001. *Helicobacter pylori* diversity within the gastric niche of a single human host. Proc. Natl. Acad. Sci. 98, 14625–14630.

Jakobsen, I.B., Easteal, S., 1996. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. Comput. Appl. Biosci. 12, 291–295.

Jolley, K.A., Chan, M.S., Maiden, M.C.J., 2004. mlstdbNet—distributed multi-locus sequence typing (MLST) databases. BMC Bioinformatics 5:86 (available from: http://www.biomedcentral.com/1471-2105/5/86).

Jorgensen, M., Daskalopoulos, G., Warburton, V., Mitchell, H.M., Hazell, S.L., 1996. Multiple strain colonization and metronidazole resistance in *Helicobacter pylori*-infected patients: identification from sequential and multiple biopsy specimens. J. Infect. Dis. 174, 631–635.

Kimura, M., 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. 11, 247–269.

Kuhner, M.K., Yamato, J., Felsenstein, J., 2000. Maximum likelihood estimation of recombination rates from population data. Genetics 156, 1393–1401.

Kuipers, E.J., Israel, D.A., Kusters, J.G., Gerrits, M.M., Weel, J., van Der Ende, A., van Der Hulst, R.W., Wirth, H.P., Hook-Nikanne, J., Thompson, S.A., Blaser, M.J., 2000. Quasispecies development of *Helicobacter pylori* observed in paired isolates obtained years apart from the same host. J. Infect. Dis. 181, 273–282.

Li, W.-H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA, pp. 309–334.

Maddison, W.P., Maddison, D.R., 1992. MacClade: Analysis of Phylogeny and Character Evolution. Version 3. Sinauer Associates, Sunderland, MA.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563.

Maynard-Smith, J., Smith, N.H., 1998. Detecting recombination from gene trees. Mol. Biol. Evol. 15, 590–599.

Montecucco, C., Rappuoli, R., 2001. Living dangerously: how *Helicobacter pylori* survives in the human stomach. Nat. Rev. Mol. Cell. Biol. 2, 457–466.

Morales-Espinosa, R., Castillo-Rojas, G., Gonzalez-Valencia, G., Ponce de Leon, S., Cravioto, A., Atherton, J.C., Lopez-Vidal, Y., 1999. Colonization of Mexican patients by multiple *Helicobacter pylori* strains with different *vacA* and *cagA* genotypes. J. Clin. Microbiol. 37, 3001–3004.

Nei, M., 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22, 2318–2342.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substation. Bioinformatics 14, 817–818.

Rambaut, A., Grassly, M.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

Raymond, J., Thiberg, J.M., Chevalier, C., Kalach, N., Bergeret, M., Labigne, A., Dauga, C., 2004. Genetic and transmission analysis of *Helicobacter pylori* strains within a family. Emerg. Infect. Dis. 10, 1816–1821.

Rothenbacher, D., Bode, G., Berg, G., Knayer, U., Gonser, T., Adler, G., Brenner, H., 1999. *Helicobacter pylori* among preschool children and their parents: evidence of parent–child transmission. J. Infect. Dis. 179, 398–402.

Salama, N., Guillemin, K., McDaniel, T.K., Sherlock, G., Tompkins, L., Falkow, S., 2000. A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. Proc. Natl. Acad. Sci. 97, 14668–14673.

Salaun, L., Audibert, C., Le Lay, G., Burucoa, C., Fauchere, J.-L., Picard, B., 1998. Panmictic structure of *Helicobacter pylori* demonstrated by the comparative study of six genetic markers. FEMS Microbiol. Lett. 161, 231–239.

Salminen, M.O., Carr, J.K., Burke, D.S., McCutchan, F.E., 1995. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by Bootscanning. AIDS Res. Hum. Retroviruses 11, 1423–1425.

Sawyer, S.A., 1999. GENECONV: a computer package for the statistical detection of gene conversion. Distributed by the author, Department of Mathematics, Washington University in St. Louis, available at http://www.math.wustl.edu/~sawyer.

Suerbaum, S., Smith, J.M., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dyrek, I., Achtman, M., 1998. Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. U.S.A. 95, 12619–12624.

Swofford, D.L., 1998. Paup* 4.0b4. Phylogenetic Analysis Using Parsimony (* and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic Inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics: Second Edition. Sinauer Associates, Inc., Sunderland, MA, pp. 407–514.

Tamura, K., Nei, M., 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.

Taylor, D.N., Parsonnet, J., 1995. Epidemiology and natural history of *H. pylori* infections. In: Blaser, M.J., Smith, P.F., Ravdin, J., Greenberg, H., Guerrant, R.L. (Eds.), Infections of the Gastrointestinal Tract. Raven Press, New York, NY, pp. 551–564.

Taylor, N.S., Fox, J.G., Akopyants, N.S., Berg, D.E., Thompson, N., Shames, B., Yan, L., Fontham, E., Janney, F., Hunter, F.M., Correa, P., 1995. Long-term colonization with single and multiple strains of *Helicobacter pylori* assessed by DNA fingerprinting. J. Clin. Microbiol. 33, 918–923.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Tobe, T., Sasakawa, C., Okada, N., Honma, Y., Yoshikawa, M., 1992. *vacB*, a novel chromosomal gene required for expression of virulence genes on the large plasmid of *Shigella flexneri*. J. Bacteriol. 174, 6359–6367.

Tsuda, M., Karita, M., Nakazawa, T., 1993. Genetic transformation in *Helicobacter pylori*. Microbiol. Immunol. 37, 85–89.

Wang, Y., Roos, K.P., Taylor, D.E., 1993. Transformation of *Helicobacter pylori* by chromosomal metronidazole resistance and by a plasmid with a selectable chloramphenicol resistance marker. J. Gen. Microbiol. 139, 2485–2493.

Worobey, M., 2001. A novel approach to detecting and measuring recombination: new insights into evolution in viruses, bacteria, and mitochondria. Mol. Biol. Evol. 18, 1425–1434.

Xia, X., 2000a. DAMBE: data analysis in molecular biology and evolution. Department of Ecology and Biodiversity, University of Hong Kong. http://web.hku.hk/~xxia/software/software.htm.

Xia, X., 2000b. Data Analysis in Molecular Biology and Evolution. Kluwer Academic Publishers, Boston, MA.

Xu, Q., Morgan, R.D., Roberts, R.J., Blaser, M.J., 2000. Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. Proc. Natl. Acad. Sci. U.S.A. 97, 9671–9676.

Yang, Z., Kumar, S., Nei, M., 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141, 1641–1650.